

The TeraPaths Testbed: Exploring End-to-End Network QoS

Dimitrios Katramatos, Dantong Yu, Bruce Gibbard
RHIC/ATLAS Computing Facility, Physics Department
Brookhaven National Laboratory
Upton, NY 11793
{dkat, dtyu, gibbard}@bnl.gov

Shawn McKee
Physics Department
University of Michigan
Ann Arbor, MI 48109
smckee@umich.edu

Abstract—The TeraPaths project at Brookhaven National Laboratory (BNL) investigates the combination of DiffServ-based LAN QoS with WAN MPLS tunnels in creating end-to-end (host-to-host) virtual paths with bandwidth guarantees. These virtual paths prioritize, protect, and throttle network flows in accordance with site agreements and user requests, and prevent the disruptive effects that conventional network flows can cause in one another. This paper focuses on the TeraPaths testbed, a collection of end-site subnets connected through high-performance WANs, serving the research and software development needs of the TeraPaths project. The testbed is rapidly evolving towards a multiple end-site infrastructure, dedicated to QoS networking research, and it offers unique opportunities for experimentation with minimal or no impact on regular, production networking operations.

Keywords—network; end-to-end; QoS; DiffServ; MPLS

I. INTRODUCTION

The TeraPaths project [1] - [3] researches the configuration of end-to-end virtual network paths, with bandwidth guarantees, across multiple administrative domains. The primary motivation of the project comes from the world of modern high energy and nuclear physics (RHIC [4], LHC [5], ATLAS [6], U.S. ATLAS [7], CMS [8]), where extremely large quantities of experimental and analysis data need to be transferred through high-speed networks across the globe, to be shared among scientists participating in various experiments. As the default behavior of the network is to treat all data flows equally, data flows of higher importance and/or urgency may be adversely impacted by competing flows of lesser importance. Furthermore, no one can yet accurately predict the time required for a data transfer. In such a network environment, the capability to prioritize, protect, and throttle the various data flows becomes of high importance, especially when considering co-scheduling of associated resources like storage systems and CPUs.

Providing an end-to-end path with guaranteed bandwidth to a specific data flow is a hard problem, because it requires the timely configuration of all network devices along the route between a given source and a given destination. In the general case, such a route passes through multiple administrative domains and there is no single control center able to perform the configuration of all devices involved. From the beginning of

the TeraPaths project it was evident that effective and practical solutions to the multitude of issues that were anticipated to arise could only be reached through study and experimentation on a suitable testbed infrastructure. Such issues included, for example, what QoS technology should be used within the end-site LANs, what options were available when dealing with administrative domains with no direct control possible, how could a WAN QoS route be automatically established, etc. The project owes its success to this testbed, which has gone through several evolution phases to allow the investigation of an expanding set of problems. The end product of the project, the TeraPaths (software) system, is being developed and tested on this testbed with the unique advantage of running in a real network environment, without any danger of adversely affecting regular site network operations.

II. BACKGROUND

A. View of the Network

A set of end-sites using the TeraPaths system for establishing QoS paths between them defines a TeraPaths site group. The network devices of the LAN of each such end site are under the control of a TeraPaths system instance. We follow the approach of conceptually dividing the end-to-end route between a source and a destination site within a site group, to three significant segments:

- The segment within the LAN of the source end-site, from the host where the data reside at the beginning of a transfer to the site's border router.
- The segment within the LAN of the destination end-site, from this site's border router to the host where the data will arrive.
- The segment connecting the source and destination site border routers, which may consist of multiple network segments belonging to different administrative domains.

There is a major distinction between the segments of the end-site LANs and the WAN segment connecting these LANs: the TeraPaths system does not have direct control over the network devices of the WAN route segment. For TeraPaths, the WAN segment is a "cloud" through which packets travel from source to destination. The WAN route segment may involve more than

one distinct WAN cloud. Different WAN clouds may have one or more peering points (see figure 1). The cloud representation is relevant in the sense that the WAN domain is an independent entity that can be contacted in a number of ways to make arrangements for a specific data flow. If the route passes through more WAN clouds (shown in figure 1 as a “WAN chain”), only one entity is responsible for configuring the route.

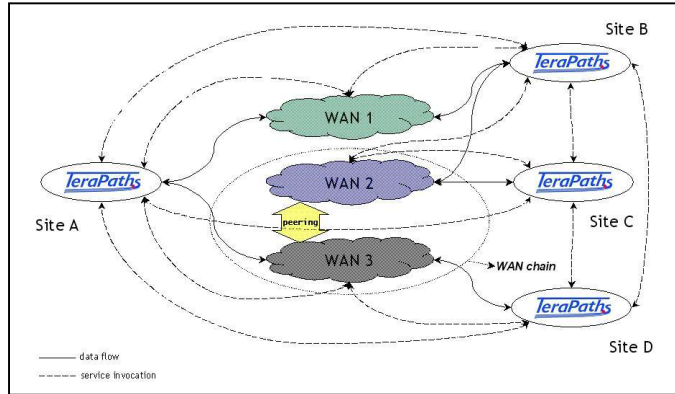


Figure 1. Conceptual view of the network.

This view of the general network reflects our experience with the realities of inter-domain cooperation. End sites are clients to WAN providers and have the capability to request special treatment for specific data flows to the degree that their provider allows. Furthermore, WAN providers are likely to adopt service level agreements between them so as to make certain QoS options available to their clients. TeraPaths follows a hybrid star/daisy chain setup model (see figure 2) where the initiating end-site coordinates with the target site and also pursues an automated WAN route setup by contacting its primary

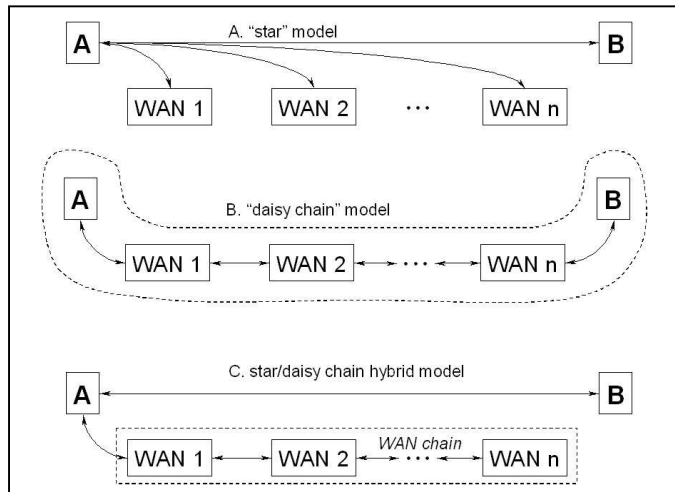


Figure 2. The three end-to-end setup models.

WAN provider and relying on that provider to coordinate, if necessary, with other WAN domains along the desired route. This approach does not require an end site to set up a route by individually contacting WAN providers along a WAN chain (star model), which would, in turn, require end-sites to have a detailed knowledge of the network so as to know which pro-

viders need to be contacted, what capabilities each one has, etc. The star model increases significantly the complexity of setting up a route and may not be always feasible. On the other hand, the daisy chain model requires all participants (end sites and WAN domains) to use a common communication protocol that allows full functionality of all basic operations of every participant (e.g., so that TeraPaths-specific parameters are guaranteed to arrive at the destination end-site). Currently, such a protocol does not exist.

B. Functionality of the TeraPaths system

The TeraPaths system is fully distributed and is implemented as a set of web service layers. Each web service layer is independent and functions as a client to lower level layers, and as a server to higher level ones. Figure 3 presents an architectural view of a TeraPaths system. Each end-site is controlled by one instance of the system software. The information required for proper system operation is strictly restricted to site-specific data only, with the exception of a set of public addresses for contacting remote TeraPaths systems and WAN providers.

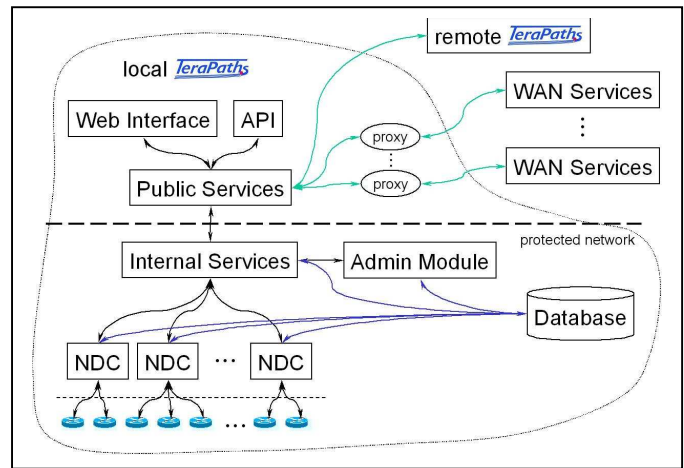


Figure 3. The TeraPaths layered web service architecture.

The TeraPaths software provides the following functionality necessary for the configuration of end-to-end network paths:

- Admission control.
- Advance reservations.
- Site LAN network device configuration (static and dynamic).
- Setup coordination with remote TeraPaths systems.
- WAN route setup arrangements with WAN providers.

TeraPaths partitions an end site’s available network bandwidth into multiple classes of service with different priorities and statically or dynamically assigned bandwidth (see figure 4). Authorized users can reserve, in advance, time slots for assigning data flows to these service classes, in essence, directing these flows through virtual network paths with bandwidth and priority guarantees. The configuration of an end-to-end path between two end sites starts with the initiating site’s TeraPaths system negotiating service class and time slot with the other

site's system. If a mutual agreement is achieved, the initiating system contacts the appropriate WAN provider to make further bandwidth arrangements for the WAN segment. The configuration of the path is successful if all three parties are successful in configuring their segment of the path.

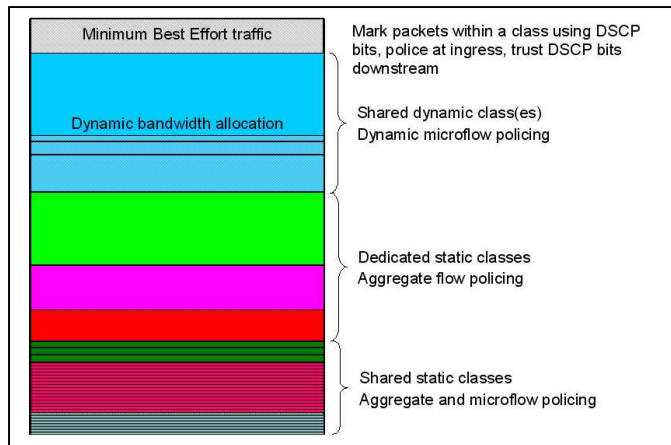


Figure 4. Bandwidth partitioning scheme.

III. NETWORK DEVICE CONFIGURATION

The creation of a virtual end-to-end QoS path depends on the capabilities of all network devices along the route, from source to destination. In this section, we discuss the choices made for TeraPaths-controlled LANs and the reasoning behind them. We also discuss methods of handling the WAN segment configuration.

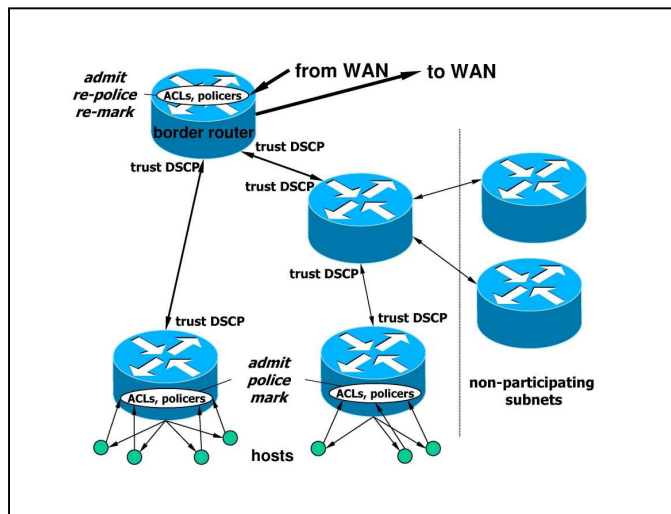


Figure 5. General example of LAN configuration.

A. End-site LAN configuration

TeraPaths uses the Differentiated Services (DiffServ) architecture [9] for configuring QoS paths within a site's LAN. The primary reasons for choosing this architecture over the Integrated Services (IntServ) and MPLS architectures [10] - [12] are its advantages in scalability and equipment compatibility.

IntServ and MPLS require all network devices along a path to maintain state information for every single data flow that receives non-default treatment. With DiffServ, only the device at the network perimeter where the privileged data flow enters needs to be configured for that flow, while the rest of the LAN is set to honor any treatment specified by the entry device. The DiffServ architecture is packet-oriented (in contrast, the IntServ and MPLS architectures are flow-oriented). Packets receive differential treatment according to their Type of Service (ToS) header markings. DiffServ utilizes six bits of a header's ToS byte for the Differentiated Service Code Point (DSCP) markings for a total of 64 different markings. The TeraPaths software can utilize all 64 resulting classes of service but falls back to the eight IP precedence classes (3-bit markings) when older equipment is involved. Part of the end site coordination is to decide which class of service to utilize if the configuration of the two sites differs. In such a case, the source and/or the destination site will need to utilize classes with compatible configuration or map the data flow to a different class.

Figure 5 shows a general example of a site LAN QoS configuration. Part of the configuration is static and needs to be included in the running configuration of all participating network devices before the TeraPaths software can have any effect. The static part of the configuration does the following:

- Enables QoS operations in all participating network devices.
- Creates the necessary QoS policies for assigning data flows to service classes.
- Assigns Access Control Lists (ACLs) to policies for admission control.
- Creates the necessary bandwidth policers for the service classes with pre-determined maximum bandwidth.
- Assigns QoS policies to device interfaces:
 - Host interface policies control (admit, police, and mark) outgoing traffic.
 - Interconnecting interface policies trust DSCP markings.
 - Border interface policies control incoming traffic (admit, and if required, re-police and re-mark).

The administration module of the TeraPaths system (currently under development) facilitates the static configuration of a site's devices.

The dynamic part of the configuration is the responsibility of the TeraPaths Network Device Controller (NDC) modules and only affects the devices on the LAN's perimeter, specifically:

- The contents (data flow definitions) of ACLs, and some service class mapping and policer parameters of the host routers.
- A similar set of parameters at the site's border router, depending on the level of trust and compatibility of site configurations.

The host routers play the role of control valves, throttling and marking data flows outgoing from hosts and incoming to the site's network. The rest of the network simply honors the packet markings all the way out of the site and into the WAN. The border router is the ultimate controller for incoming traffic: unless all incoming traffic from a specific external source or site is trusted, the border router admits data flows into local service classes through ACLs. Additionally, the border router may re-police an incoming flow to ensure that the agreed upon bandwidth limits are observed, and/or change the markings of said flow and assign it to a different service class supported by the local site configuration.

The device configuration steps for admitting a data flow in a service class differ according to the kind of service assigned by TeraPaths to this class. A class with statically allocated bandwidth (pre-determined bandwidth) only requires admission control by adding suitable rules to the necessary ACLs. A class with dynamically allocated bandwidth requires, except for admission control, the dynamic configuration of a policer dedicated to a specific data flow (or set of flows), while the TeraPaths system keeps track of the assigned utilization of bandwidth for this class to ensure there is no oversubscription of resources.

B. WAN configuration

Configuring the network segment between two end sites is an operation requiring the existence of service level agreements of said end sites with one or more WAN providers, and between the WAN providers themselves. The automated configuration of the WAN segments is more complex, requiring the existence of suitable software mechanisms exposed by the WAN providers and collaboration between the responsible parties. The level of QoS that is possible within a WAN segment affects the level of QoS guarantees that TeraPaths can provide. Typical cases are the following:

1) No WAN QoS available:

Only overall throttling of a data flow is possible. Prioritization/protection is possible only within end site LANs, while the flow will be treated as standard, "best effort" traffic in the WAN. If the WAN becomes increasingly loaded, the requested bandwidth may not be honored due to lack of QoS.

2) WAN QoS only statically available:

End-to-end data flow prioritization/protection and throttling is possible, however, the WAN providers along a route have to agree to configure in advance their network devices using Diff-Serv techniques similar to those at the end sites or MPLS. This kind of WAN configuration is essentially a long-term Service Level Agreement (SLA), which becomes effective only in the presence of privileged data flows. Thus, while the end site LANs can establish a virtual path on demand, the WAN segment is in "stand-by" mode.

3) WAN MPLS tunnel dynamic configuration:

End-to-end prioritization/protection and throttling is possible on demand. This case assumes that WAN providers have publicly available services allowing the dynamic configuration of MPLS tunnels between specific end-sites. A user is able to set up a virtual path with flow-level granularity (defined by two

pairs consisting of a single IP address and a single port number).

In an ideal situation, the services of a chain of WANs can coordinate for establishing an MPLS tunnel across WAN domains, so that an end site needs to invoke the services of only one WAN provider. It is also possible to encounter WAN segments configured both statically and dynamically along a desired route.

IV. THE TESTBED

A major advantage offered by a dedicated network testbed is the ability to experiment and study different ways of configuring the network devices in order to achieve the creation of the desired virtual paths, and all without affecting the conventional user of the network. Several solutions and techniques adopted by TeraPaths are the result of multiple iterations of research, development, and experimentation on the TeraPaths testbed infrastructure. The testbed went through a number of evolution phases, each one corresponding to a constantly evolving research focus.

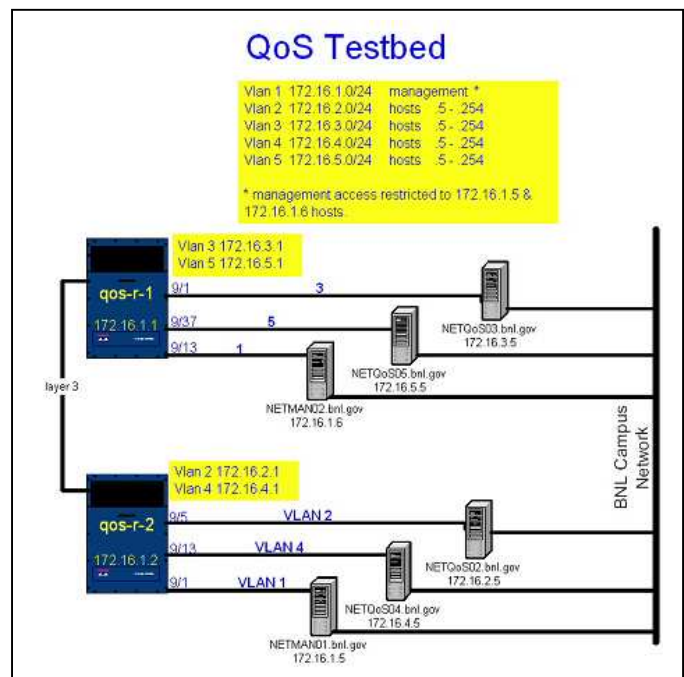


Figure 6. The initial TeraPaths testbed.

A. Simulation phase: 2-site simulation, private network

The first version of the testbed was local to BNL and was comprised of two Cisco 6500 series routers (switches with SUP2 supervisor engine and routing capabilities) and six multi-homed hosts connected with a private network (see figure 6). Similar hardware is used in the actual production network at BNL. The two routers were connected with a 1 Gbps fiber line. Each router had three hosts connected, one of which was the device management node. The management node was the only node allowed to access the specific router and perform configuration changes. The other two hosts were utilized for creating

regular and prioritized traffic using the corresponding hosts of the other router as counterparts.

In this version, the testbed was utilized in two stages:

- Initially, for trying the operation of various Cisco IOS QoS-related commands, and for the development and testing of the initial versions of the TeraPaths software, especially for making decisions about the core functionality of the device controller layer. With the help of the testbed, we defined an interface for a general QoS router device, independent of hardware type, and proceeded with an implementation of this interface for Cisco hardware. Support for different hardware is possible with suitable device drivers, as long as this hardware can in turn support DiffServ, fully or partially.
- Subsequently (after reconfiguring routers and software) the testbed was used for simulating end-to-end setups between two sites, and verifying the more complex device configurations necessary in such setups. In this testbed version, the 1 Gbps connection between the two routers corresponds to a dedicated MPLS tunnel of equal bandwidth.

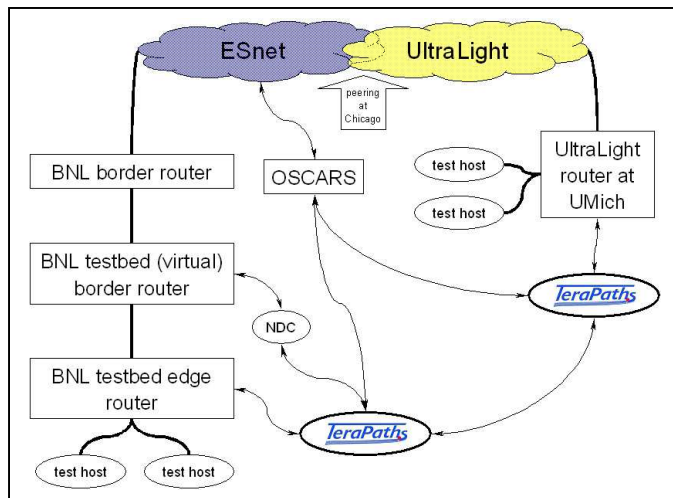


Figure 7. The current version of the TeraPaths testbed.

B. Current phase: 2-site installation, public network

This phase of the testbed involves two TeraPaths end-sites, one at BNL and one at the University of Michigan (UMich), connected through the ESnet [13] and UltraLight [14] networks (see figure 7), which have a peering point at Chicago. WAN MPLS tunnel requests are directed to ESnet's OSCARS service [15]. The original BNL testbed was modified to represent a single end-site (see figure 8), while a second end-site was put together at the University of Michigan. The BNL border router was set to trust the traffic from the testbed, while the original testbed's second router was set to play the role of the border router and thus accept the necessary configuration commands from its NDC module. The configuration of the virtual border router is identical to that needed for the actual border router; however, possible errors encountered during

testing cannot affect the actual border router, which is critical for regular site operations.

The UltraLight network does not expose a service equivalent to OSCARS. Thus, no automated MPLS tunnel configuration is currently possible. To allow packet DSCP markings to move freely between BNL and UMich, we chose to statically configure all devices between the end of the OSCARS-configured MPLS tunnel at Chicago and the UltraLight router

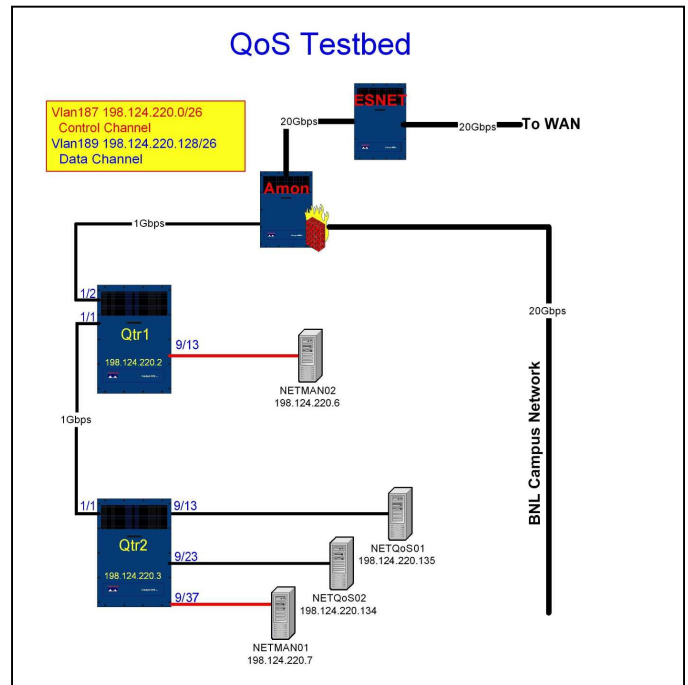


Figure 8. The BNL-side part of the current TeraPaths testbed.

at UMich to trust the markings of traffic between the involved subnets at BNL and UMich (see figure 9). Thus, we have a combination of the WAN configuration cases presented in sections III.B.2 and III.B.3 above.

The UltraLight router at Chicago is considered both border and host router for the UMich TeraPaths site, i.e., this router's role is to control QoS-wise both outgoing and incoming data traffic. The static configuration of the devices between the end of the MPLS tunnel and this router ensures that the intended prioritization of a data flow will be respected along a segment of a route that doesn't have automated QoS support.

C. Expanding phase: multi-site installation, public network

Our plans for the next evolution phase of the TeraPaths testbed are to expand it to a multiple end-site installation. This will be done in the context of a set of planned "Network Challenges" for LHC, scheduled for March-April of 2007. As part of these challenges, various Tier-2 centers participating in the ATLAS and CMS experiments will deploy network manage-

ment and control applications that are being developed in various NSF and DOE-funded network research projects. We anticipate installing TeraPaths at three U.S. ATLAS Tier-2 sites: University of Chicago/Indiana University, University of Oklahoma/University of Texas at Arlington, and Stanford Linear Accelerator Center (SLAC). The expanded testbed will

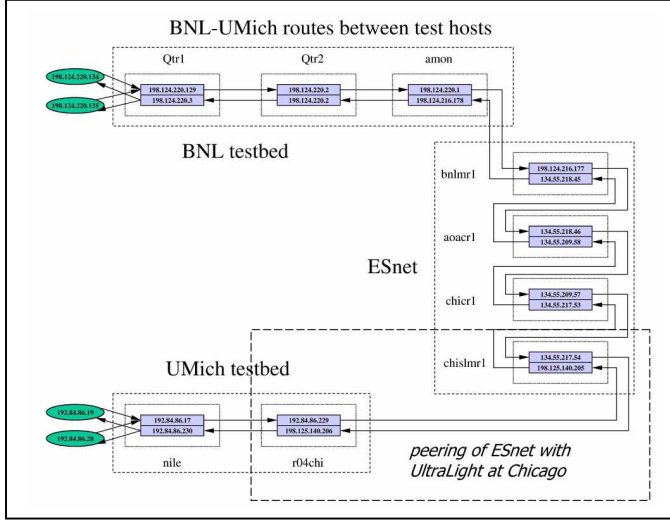


Figure 9. TeraPaths testbed routes.

allow us to verify the correctness of new operations and evaluate the system’s functionality, scalability, and response time in an almost-production-grade environment with multiple users and multiple sites. Such an environment will also test TeraPaths features like support of unidirectional and bidirectional QoS for flows (meant for increasing the utilization of service classes) and facilitate experimental measurements on the effectiveness of QoS in a data-intensive environment and its impact on conventional data traffic.

V. LESSONS LEARNED

The experience of using a dedicated testbed has helped us learn a number of valuable lessons that have affected, and will continue to affect our research directions and system design decisions. A summary of lessons learned so far is the following:

- Flow policing within a router comes at a cost. We have observed throttled bandwidths to be typically 2 to 5%, and up to 10%, less than the corresponding policer configuration. One needs to compensate for this cost by suitably increasing the corresponding policer’s bandwidth configuration.
- Policing adds some additional “virtual resistance” to a path, which can affect the window of the standard TCP/IP protocol. Often, multiple data streams are necessary to achieve the desired bandwidth in a transfer. The use of aggregate bandwidth policers allows multiple data flows to share a virtual path. All packets from all flows carry the same DSCP markings in this case.

- It is imperative to leave at least 10% of the available bandwidth for best effort (default) traffic to allow for conventional traffic and prevent privileged data flows from “freezing up”. When a unidirectional QoS configuration is used, the acknowledgement packets from destination to source travel as best effort. Without sufficient bandwidth allocated for best effort traffic, the flow of acknowledgements will be disrupted and will subsequently disrupt the primary flow in the opposite direction, be it privileged or not.
- While bidirectional QoS configurations guarantee the flow of acknowledgements, reserving equal amounts of bandwidth for both directions may be a waste of resources, as acknowledgment-only packets carry significantly smaller payload. There are three ways to deal with this case. In order of increasing system complexity, these ways are:
 - Allow acknowledgements travel as best effort (default TeraPaths behavior).
 - Direct all acknowledgement packets from all flows into a dedicated class of service with higher priority than best effort.
 - Use bidirectional QoS with asymmetrically reserved bandwidth to minimize resource waste.
- The system command channel is best to be permanently directed (through static configuration) into a high-priority class of service to guarantee timely cooperation of end sites in a congested environment.

VI. CONCLUSIONS

The TeraPaths project demonstrates that the combination of LAN QoS techniques, based on the DiffServ architecture combined with WAN MPLS tunnels, is a feasible and reliable approach to providing end-to-end, dedicated bandwidth paths to data flows in a demanding, distributed, computational environment, such as the environment needed for high energy and nuclear physics research. TeraPaths technology offers a flexible way to partition a site’s available bandwidth into pre-determined bandwidth slots, and to protect various data flows from competing against each other. Critically important to the research and development efforts has been, and continues to be, the unique testbed infrastructure that is available to the project. This infrastructure has allowed us to conduct research that would have been otherwise impossible in a production environment, due to the inherent danger of disruption of operations.

This unique testbed helped the TeraPaths project, along with our close collaboration with ESnet’s OSCARS team, to reach an important milestone in the summer of 2006, when the world’s first end-to-end, fully-automated QoS path setup took place between BNL and UMICH.

With the further planned expansion of the infrastructure, we are looking forward to continuing our research and addressing new QoS networking problems. We will continue the development of the software system to achieve production quality and

robustness, with the intent to fully utilize TeraPaths in production within the ATLAS framework and elsewhere.

REFERENCES

- [1] (2007, Feb.). The TeraPaths End-to-End QoS Networking Project. [Online]. Available: <http://www.racf.bnl.gov/terapaths>
- [2] S. Bradley, F. Burstein, L. Cottrell, B. Gibbard, D. Katramatos, Y. Li, S. McKee, R. Popescu, D. Stampf, D. Yu. "TeraPaths: a QoS-enabled collaborative data sharing infrastructure for peta-scale computing research." Proceedings of Computing in High Energy and Nuclear Physics (CHEP 2006), T.I.F.R., Mumbai, India, Feb. 13-17, 2006.
- [3] D. Katramatos, B. Gibbard, D. Yu, S. McKee. "TeraPaths: end-to-end network path QoS configuration using cross-domain reservation negotiation." Proceedings of the 3rd International Conference on Broadband Communications, Networks, and Systems (BROADNETS 2006), San Jose, California, Oct. 1-5, 2006.
- [4] (2007, Jan.). Relativistic Heavy Ion Collider, RHIC. [Online]. Available: <http://www.bnl.gov/RHIC/>
- [5] (2007, Jan.). The Large Hadron Collider (LHC). [Online]. Available: <http://lhc.web.cern.ch/lhc/>
- [6] (2007, Jan.). The ATLAS experiment. [Online]. Available: <http://atlas.ch/>
- [7] (2007, Jan.). The U.S. ATLAS project. [Online]. Available: <http://www.usatlas.bnl.gov/>
- [8] (2007, Jan.). The CMS experiment. [Online]. Available: <http://cms.cern.ch/>
- [9] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, W. Weiss. (1998, Dec.). An architecture for differentiated services. IETF RFC 2475. [Online]. Available: <http://www.ietf.org/rfc/rfc2475.txt>
- [10] R. Braden, D. Clark, S. Shenker. (1994, June). Integrated Services in the Internet Architecture: an Overview. IETF RFC 1633. [Online]. Available: <http://www.ietf.org/rfc/rfc1633.txt>
- [11] R. Braden, L. Zhang, S. Berson, S. Herzog, S. Jamin. (1997, Sep.). Resource ReSerVation Protocol (RSVP). IETF RFC 2205. [Online]. Available: <http://www.ietf.org/rfc/rfc2205.txt>
- [12] E. Rosen, A. Viswanathan, R. Callon. (2001, Jan.). Multiprotocol label switching architecture. IETF RFC 3031. [Online]. Available: <http://www.ietf.org/rfc/rfc3031.txt>
- [13] (2006, Dec.). The Energy Sciences Network. [Online]. Available: <http://www.es.net/>
- [14] (2007, Jan.). UltraLight: An Ultrascale Information System for Data Intensive Research. [Online]. Available: <http://www.ultralight.org/>
- [15] (2007, Jan.). ESnet On-demand Secure Circuits and Advance Reservation System (OSCARS). [Online]. Available: <http://www.es.net/oscars/>